Design document

The implementation is divided into the following steps/jobs.

1. Count the number of nodes in the graph.
2. Get the number of edges in the graph, calculate minimum, maximum and average number of edges per node.
3. Prepare initial file with the format < NODE ID: updated rank: last rank :<adjacency list>>
   Initial rank is given as (1/total nodes), in the beginning both updated and last rank are set to this value.
4. Specify maximum number of iterations and slackness which defines the accuracy of convergence.
5. Calculate the contribution from dangling nodes.
6. Update the page-rank –
   Uses the file from previous iteration as input, takes into account dangling nodes.
   Makes use of the PageRank formula specified in paper with damping factor = .85.
7. Check if the ranks have converged.
8. Iterate steps 5,6,7 until convergence is reached or the maximum iteration limit is breached.

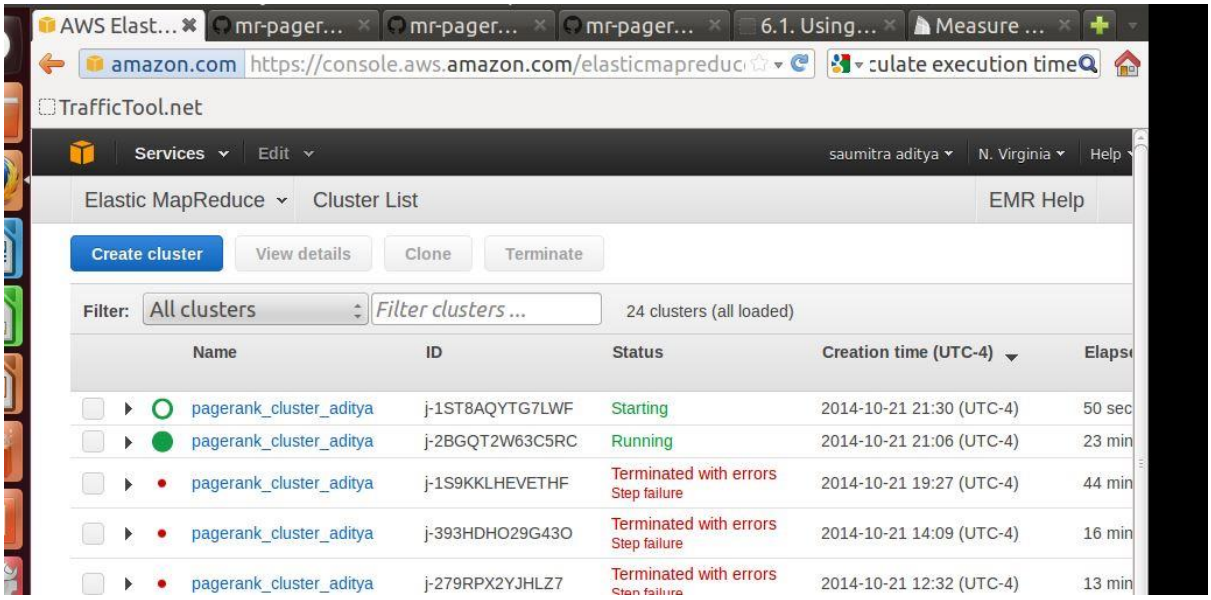
Method for checking convergence:

I calculate the difference between the updated and last pagerank of a node and sum it over all the nodes, I call this value total_delta. As one of the arguments I specify a slackness factor eg.01 i.e. I will iterate until sum of delta over all nodes is not less than .01.

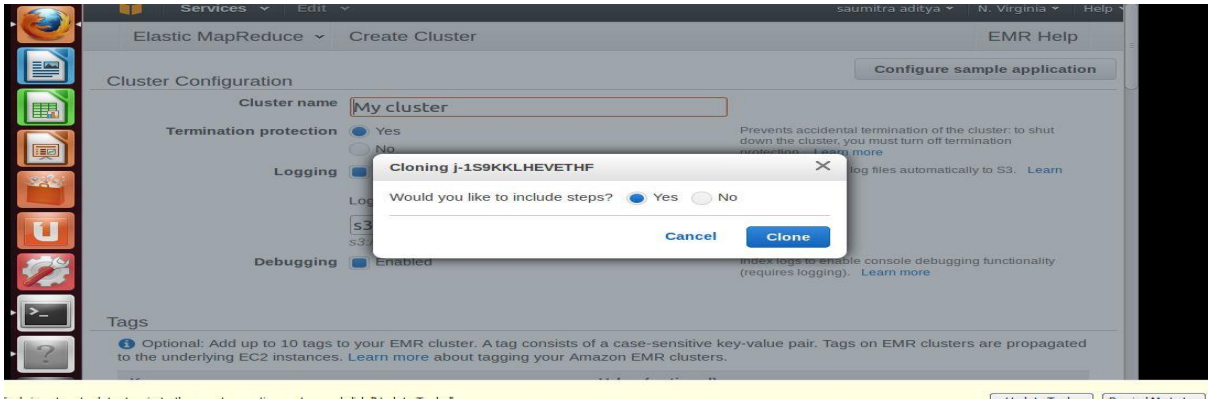
Handling dangling nodes

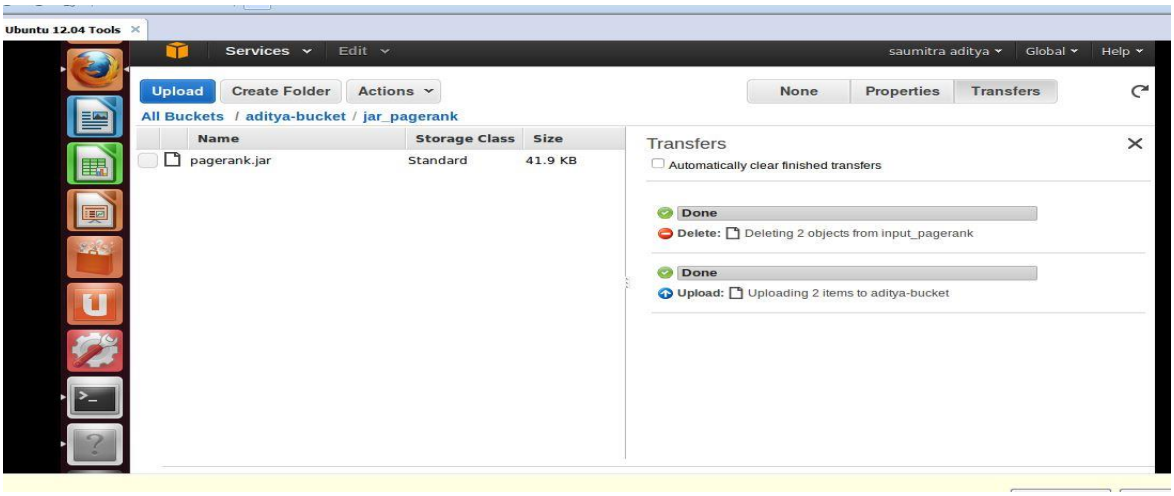
The pagerank of dangling node is distributed among all the nodes in equal proportions i.e every node in the graph gets (pagerank of dangling node/ total nodes) contribution from the dangling node.

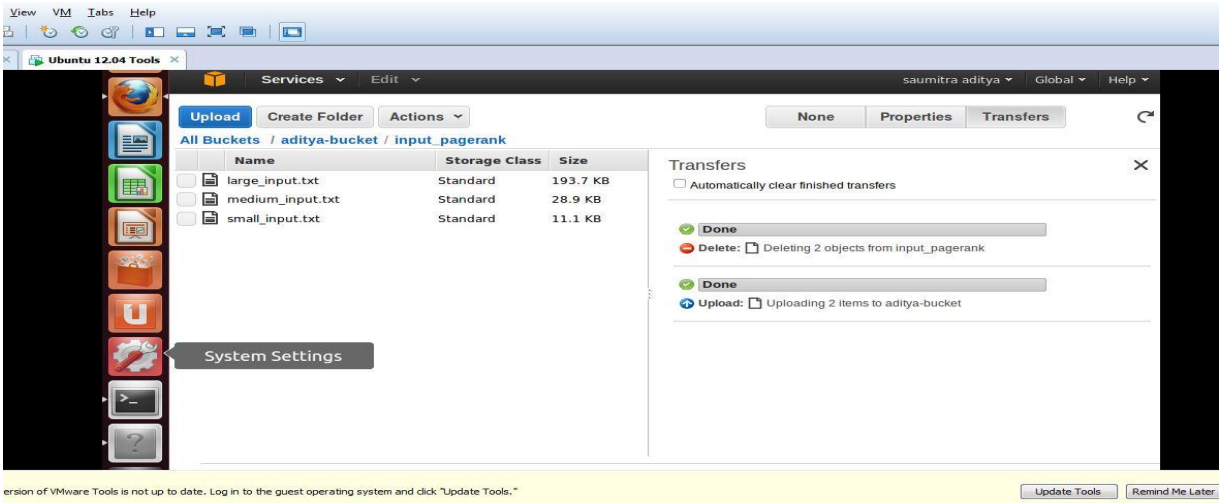
Steps on AWS. c

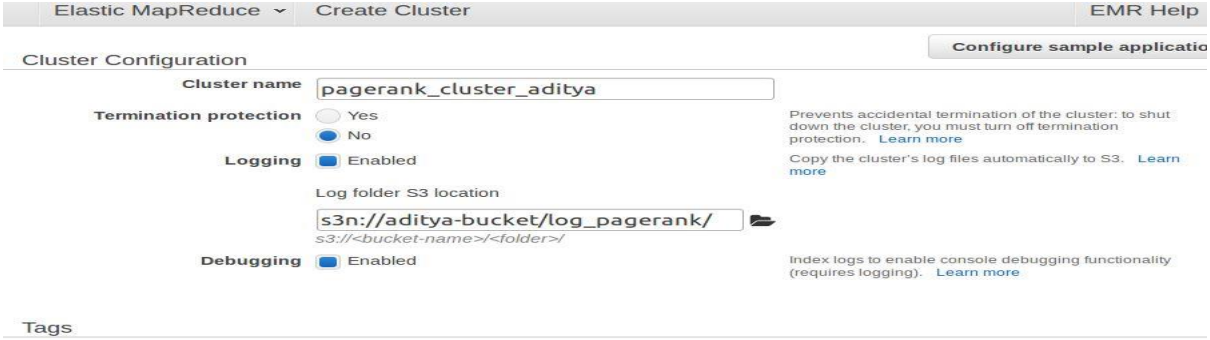


Clone a cluster, upload jar file and input files.

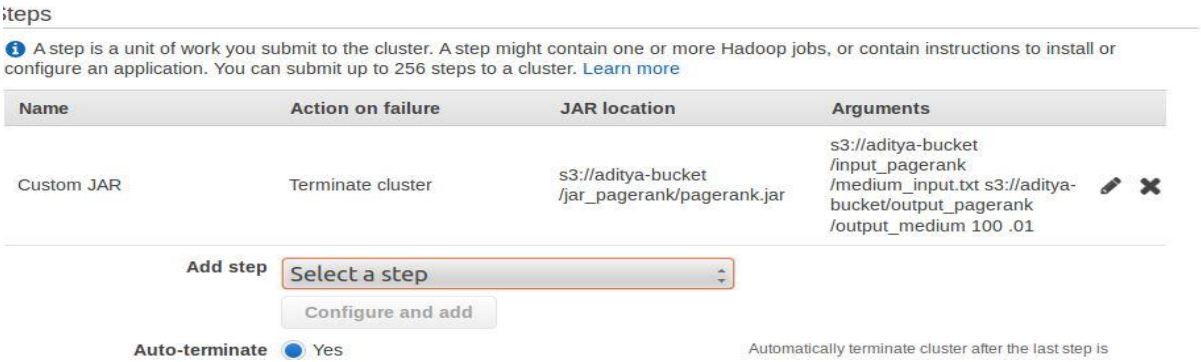

Services
Edit
saumitra aditya
N. Virginia
Help
Elastic MapReduce
Create Cluster
EMR Help
Configure sample application
Cluster Configuration
Cluster name
My cluster
Termination protection
Yes
No
Prevents accidental termination of the cluster; to shut down the cluster, you must turn off termination protection. Learn more
Logging
log files automatically to S3. Learn
Cloning j-1S9KKLHEVETHF
Would you like to include steps?
Yes
No
Cancel
Clone
Debugging
Enabled
index logs to enable console debugging functionality (requires logging). Learn more
Tags
Optional: Add up to 10 tags to your EMR cluster. A tag consists of a case-sensitive key-value pair. Tags on EMR clusters are propagated to the underlying EC2 instances. Learn more about tagging your Amazon EMR clusters.


Ubuntu 12.04 Tools
Services
Edit
saumitra aditya
Global
Help
Upload
Create Folder
Actions
None
Properties
Transfers
All Buckets / aditya-bucket / jar_pagerank
Name
Storage Class
Size
pagerank.jar
Standard
41.9 KB
Transfers
Automatically clear finished transfers
Done
Delete: Deleting 2 objects from input_pagerank
Done
Upload: Uploading 2 items to aditya-bucket

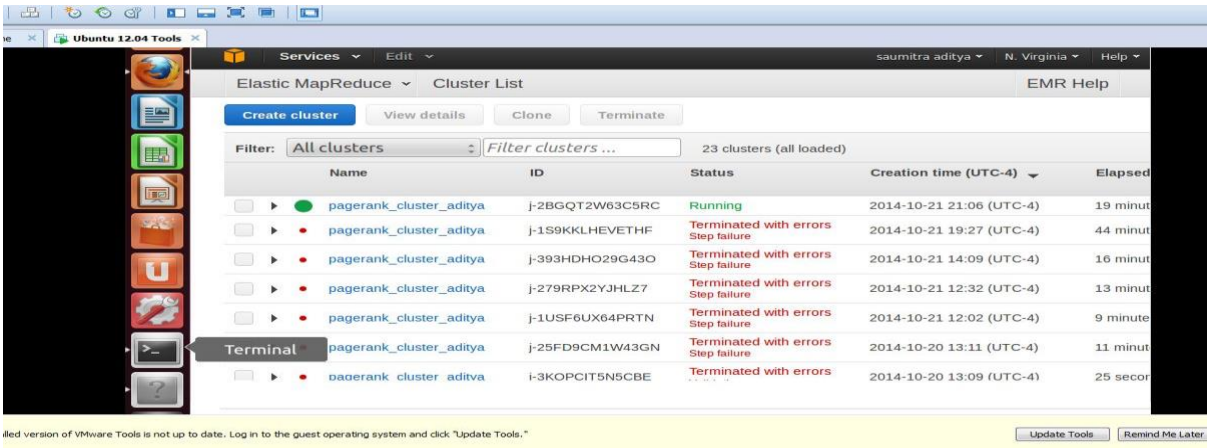

View
VM
Tabs
Help
Ubuntu 12.04 Tools
Services
Edit
saumitra aditya
Global
Help
Upload
Create Folder
Actions
None
Properties
Transfers
All Buckets / aditya-bucket / input_pagerank
Name
Storage Class
Size
large_input.txt
Standard
193.7 KB
medium_input.txt
Standard
28.9 KB
small_input.txt
Standard
11.1 KB
Transfers
Automatically clear finished transfers
Done
Delete: Deleting 2 objects from input_pagerank
Done
Upload: Uploading 2 items to aditya-bucket
System Settings
ersion of VMware Tools is not up to date. Log in to the guest operating system and click "Update Tools."
Update Tools
Remind Me Later



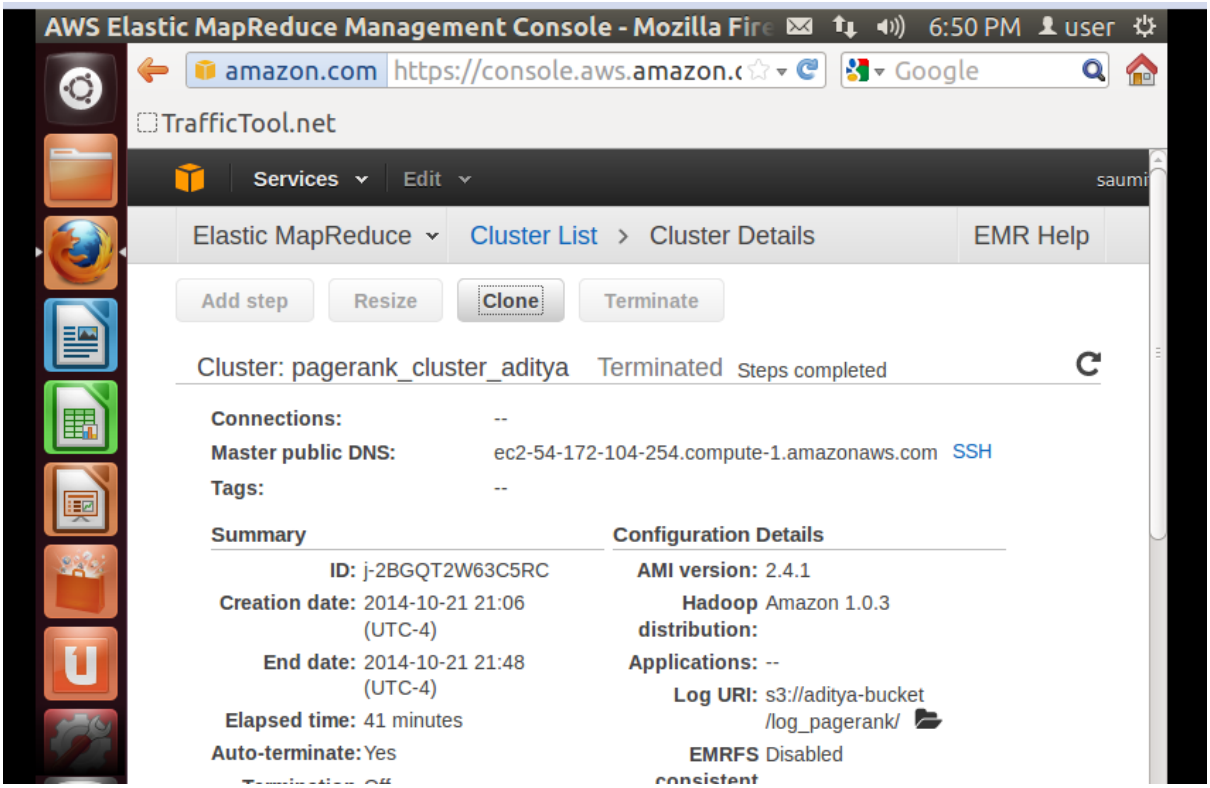
Set location of log files.



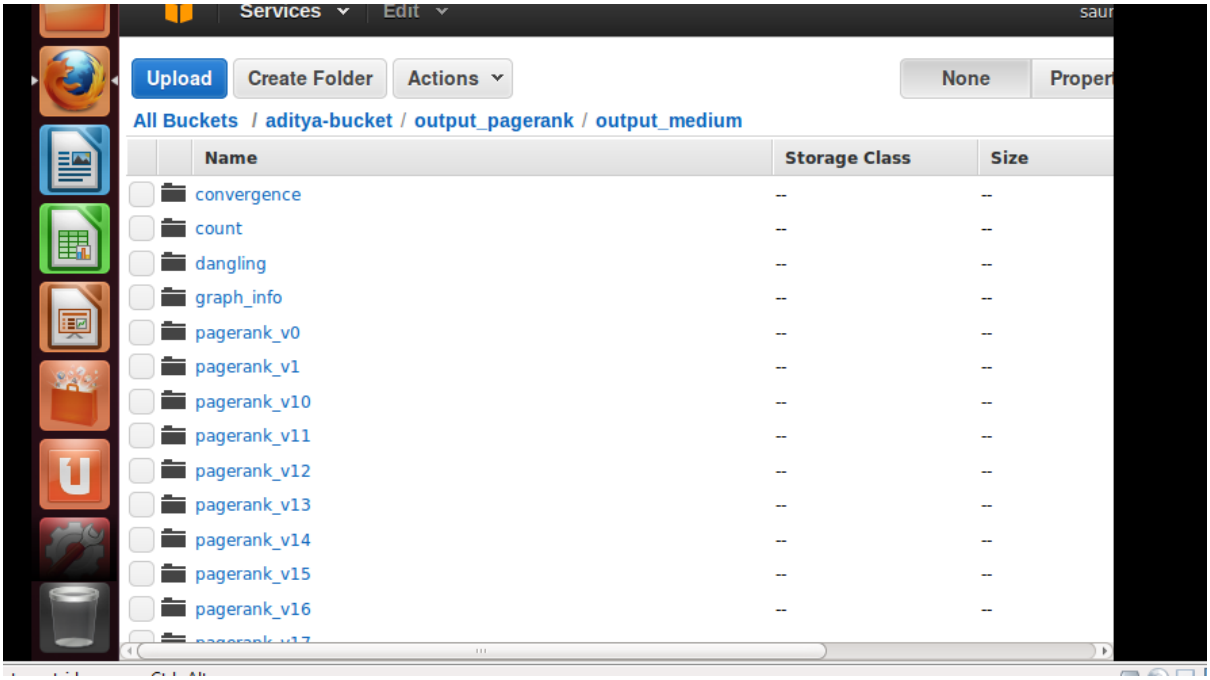
Specify location of custom jar on S3 and the arguments for the job like input path, output path, maximum iterations and slackness limit.



Start the cluster.



Wait for job to complete.



Output files.

Results.

Small input.

graph_summary total_nodes 93 total_links 195 min_links 0 max_links 5 avg_links 2.0

Top ten nodes

1045242120554427576659079652435600451 01 0.0032298502535568416
8230615676619458762969035008396747339 4 0.002391805989823018
1346619962341598084883751700704731875 82 0.0023640307765620604
1553466171085608085811841426293297292 30 0.002333236338016026
2271094487023021135079036047599184160 63 0.002242148643274156
1686732524695505795578990678364209325 46 0.002120614257327908
5797937061574192860928300503432522008 8 0.0018038879600958434
1764959878348252574507371016761860610 7 0.0016266018873613973
9766802053812480883806535626787288787 1 0.0015550908078008105
2842077003105952850518496645604738149 89 0.001242114789023274

Metrics
convergence_count= 17
pagerank job completed in 2399015ms

https://s3.amazonaws.com/aditya-bucket/output_pagerank/output_small/sorted_output/part-r-00000

https://s3.amazonaws.com/aditya-bucket/output_pagerank/output_small/graph_info/part-r-00000

Medium input

graph_summary total_nodes 316 total_links 430 min_links 0 max_links 5 avg_links 1.0

Top ten nodes

1119814434226675999161016412674149708 74 0.0013445596494086798
3044267606251528441559872341801435506 1 7.510357058956637E-4
2171823983447171219850599123458539983 16 6.580261564678982E-4
6436328214894587621089033687286575534 3 5.074033228553838E-4
1041058446974700132763723317838940767 26 4.92268950932983E-4
2551412718718875726042049542077692795 63 4.900699868634195E-4
1164807726293620120024606267770816054 00 4.73769838798419E-4
2986907431350775008020078516080464389 95 4.516332036552296E-4
3038068320535662905720957163526499816 43 4.390352259543741E-4
Metrv1485118383610641044116536732264840 3910 4.37235137198252E-4

Metrics

convergence_count= 17
pagerank job completed in 2246659ms

https://s3.amazonaws.com/aditya-bucket/output_pagerank/output_medium2/sorted_output/part-r-00000

https://s3.amazonaws.com/aditya-bucket/output_pagerank/output_medium2/graph_info/part-r-00000

Large_input.txt

graph_summary total_nodes 1458 total_links 3545 min_links 0 max_links 5 avg_links 2.0

Top ten nodes

640329419637502236015056967871231384 45 9.969838542203818E-4

119337412437940133144881923208049882442 8.582750615375713E-4

294418289840301973322672169300394924184 6.918287599413247E-4

171296782295871349005571652832417803 6 6.388412220270341E-4

284510239251910046427593057486449185085 5.848073863619843E-4

214917686594559236497547622533457258166 5.628633041528273E-4

131802765080162628666440881374317948788 5.56646529785805E-4

156471617313644419826686265789184402299 5.44236473830064E-4

269011742891273932588164655856253091447 5.358855817756014E-4

259479793941959149960309933682866059975 5.344815197232751E-4

Metrics

convergence_count= 17

pagerank job completed in 2219340ms

https://s3.amazonaws.com/aditya-bucket/output_pagerank/output_large2/sorted_output/part-r-00000

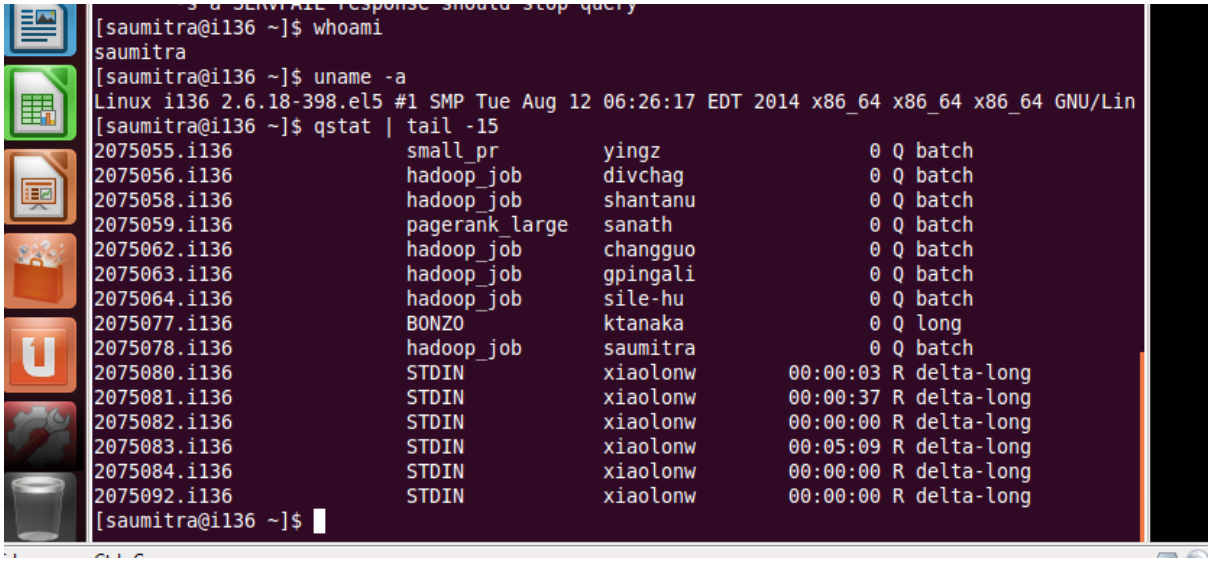
https://s3.amazonaws.com/aditya-bucket/output_pagerank/output_large2/graph_info/part-r-00000

Conclusion:

The metrics indicate that all the three graphs converge after 17 iterations when slackness factor is .01 (it is the sum of delta of latest and last PR over all nodes.)

The timing metric from AWS is not accurate, while computing on local machine larger the input size more is the running time.

I was not able to run jobs on Future Grid as it is still in queue after waiting for more than a day.



```
[saumitra@i136 ~]$ whoami
saumitra
[saumitra@i136 ~]$ uname -a
Linux i136 2.6.18-398.el5 #1 SMP Tue Aug 12 06:26:17 EDT 2014 x86_64 x86_64 x86_64 GNU/Lin
[saumitra@i136 ~]$ qstat | tail -15
2075055.i136                small_pr         yingz                  0 Q batch
2075056.i136                hadoop_job       divchag                0 Q batch
2075058.i136                hadoop_job       shantanu               0 Q batch
2075059.i136                pagerank_large   sanath                 0 Q batch
2075062.i136                hadoop_job       changguo               0 Q batch
2075063.i136                hadoop_job       gpingali               0 Q batch
2075064.i136                hadoop_job       sile-hu                0 Q batch
2075077.i136                BONZO            ktanaka                0 Q long
2075078.i136                hadoop_job       saumitra               0 Q batch
2075080.i136                STDIN            xiaolonw        00:00:03 R delta-long
2075081.i136                STDIN            xiaolonw        00:00:37 R delta-long
2075082.i136                STDIN            xiaolonw        00:00:00 R delta-long
2075083.i136                STDIN            xiaolonw        00:05:09 R delta-long
2075084.i136                STDIN            xiaolonw        00:00:00 R delta-long
2075092.i136                STDIN            xiaolonw        00:00:00 R delta-long
[saumitra@i136 ~]$
```